Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., author comment AC1
https://doi.org/10.5194/hess-2021-392-AC1, 2021

Principle 2 Section Inclusion for hess-2021-392

Caitlyn A. Hall et al.

Author comment on "A hydrologist's guide to open science" by Caitlyn A. Hall et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-392-AC1, 2021

Upon pre-print publication, the following section was omitted. Attached is the preprint with it included in line with the rest of the text.

Principle 2 – Open Data: Open hydrologists document all components of their data collection and analysis pipeline, favoring open and non-proprietary technologies.

Hydrologists often combine data from a wide variety of field, laboratory, and computer sources, such as streamflow gauges, water samples, remote sensing datasets, digital elevation models, land use maps, and meteorological data. Data quality can only be assessed, and potential results replicated when the hardware design and specifications of measurement tools and data loggers are available to the public. We encourage use of open (i.e. non-proprietary) data formats, hardware specifications and software in data collection and processing workflows, and their systematic documentation with the aim to enable their re-use by the interested reader.

Data from the laboratory is often exported in formats specific to the laboratory device and typically requires some data reformatting necessary for post-processing. The format of computer-generated data (e.g., hydrology model outputs) varies with the computer software that generated it. An open data collection and analysis pipeline includes information on (1) hardware and software used, (2) original and processed (meta-)data and databases, (3) data processing and analysis techniques and tools used, and (4) documentation of the overall analysis process, including assumptions and perceptual models (see Principle 1). Re-usability and transferability of software and data processing pipelines greatly accelerates scientific progress in hydrology by reducing time wasted on re-inventing the wheel, helping discover problems in the analysis, and improving the quality of hydrologic research.

Practical Guide to Open Data Collection and Analysis

Open hydrologists share and cite the source and collection method of all qualitative and quantitative data involved in their research, including field, laboratory, computer, and/or third-party (online) data used. A current list of data repositories commonly used by hydrologists that adhere to open science standards is kept on open-hydrology.github.io. The best place to store data for an open hydrology project depends on the type and size

of the data, the specific scientific domain, and other requirements stipulated by the funders and stakeholders. If an open hydrology study relies on third-party data that is not (yet) open, ask the original data creators to make the data or a data subset publicly available. Archived original, intermediate, and final versions of all data used to obtain the results of a particular study are crucial for reproducing open hydrology research. See Principle 4 for more details on publishing data.

To make data and analysis sharing more straightforward, a data management plan should be developed in the early stages of the research project, emphasizing open data principles and maintaining cyberinfrastructure and community standards. Data management plans describe where data will come from, what formats it will be stored in, who will manage and maintain it, how privacy will be maintained (if applicable), and how data and results will be shared and stored in the short- and long-term. Data management plans may be required by funders where they are typically limited in length. However, extended data management plans can increase research project transparency, and can be created using publicly available templates (e.g., ckan, DMPTool, resources.data.gov) that adhere to funder requirements and formatting. Some tools (e.g., ckan) can help hydrologists make previously unpublished data publicly available, even after publication.

Open hydrologists should explicitly provide public access (e.g., through a link accessible on the journal publication site) to: (1) raw data and associated metadata (including specifications of the devices used to collect data), (2) descriptions and citations for the analysis methods and software versions used, (3) workflows, code, and software developed to collect and analyze data, (4) descriptions of quality controls used when processing raw data, (5) final processed data, and (6) descriptive methods used to integrate data into other processing tools. The level of detail necessary to ensure openness can differ wildly between studies. When data sources, processing, and accessibility are complex, additional descriptions in an appendix or supplementary information may be appropriate upon publication of hydrologic research.

Ideally, all data used to draw conclusions should be published publicly to facilitate reproducibility, but copyright on third-party data, privacy, or other issues related to data sensitivity may prohibit open publication of all underlying data. Discuss, agree, and document with your collaborators what can be shared publicly as early as possible. If certain datasets cannot be shared publicly, add a statement to the final publication explaining what conditions need to be fulfilled to obtain access to the data and why some data remain private. Relevant resources and local guidelines for data anonymization and sharing (e.g., General Data Protection Regulation) need to be considered before developing a data management plan and conducting research (Zipper et al., 2019). When making data publicly available, open hydrologists strive to store data in universal, non-proprietary, and software agnostic formats that are compatible with most operating systems and include metadata (data about the data that provides background context). For example, text and tabulated data can be stored as standard American Standard Code for Information Exchange (ASCII) text (American Standard Code for Information Interchange) instead of proprietary or software-specific types (e.g., Microsoft Word .docx or Excel .xlxs files) that require a paid software license to use. Even if it might be computationally efficient, avoid creating new file types that are specific to a certain model or software. For most hydrologic data, NetCDF (i.e., .nc) files are currently the gold standard for storing data and metadata. If metadata cannot be part of the data (file) itself, store the metadata in as close proximity to the data as possible. For example, open hydrologists can include links in the metadata to where the data is stored and vice versa. They can also use standard naming and unit conventions (e.g., SI units), metadata formats (e.g., Water Metadata Language), and be informative and sufficiently complete to allow for a better understanding of the data and reproduction of study results.